Conceitos Básicos de MD

- MD extrai modelos a partir de dados observados
- Modelos representam o conhecimento induzido
- Análise de modelo por ser humano
  - Subjetivo
  - Avalia se os modelos trazem conhecimento útil ou interessante

Conceitos Básicos de MD

- MD lida com dados de observações, não dados experimentais
  - Dados que foram coletados para um propósito diferente de análise por MD
    - Ex. Dados coletados para atualizar registros de pacientes de um hospital
  - Objetivos da aplicação não deve influenciar a estratégia de coleta de dados
- Maioria dos métodos de MD são baseados em algoritmos de Aprendizado de Máquina (AM)

Aprendizado de Máquina

- Investiga técnicas computacionais capazes de adquirir automaticamente
  - Novas habilidades
  - Novo conhecimento
  - Novas formas de organizar o conhecimento existente
- Definição
  - Técnicas de AM podem melhorar seu desempenho em uma dada tarefa utilizando experiências prévias

Mitchell, 1997

Aplicações de AM

- Programas baseados em AM têm sido bem sucedidos para:
  - Reconhecer palavras faladas
  - Reconhecimento de faces
  - Predizer taxas de cura de pacientes de pneumonia
  - Detectar uso fraudulento de cartões de crédito
  - Analisar dados de expressão gênica
  - Prever estrutura de proteínas

Aplicações Clássicas de AM

- Aprender a reconhecer palavras faladas
  - SPHINX (Lee 1989)
- Aprender a conduzir um automóvel
  - ALVINN (Pomerleau 1989)
- Aprender a classificar objetos celestiais
  - (Fayyad et al 1995)
- Aprender a jogar gamão
  - TD-GAMMON (Tesauro 1992)

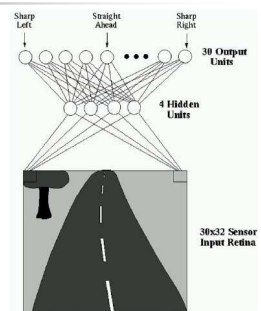
ALVINN

Dean Pomerleau
CMU

ALVINN

- Sistema automático de navegação para automóveis
  - Baseado em uma câmera montada no veículo
  - Dirigiu a 70 M/h (110 Km/h) em uma rodovia pública americana
    - De costa a costa em 1989 por 2850 milhas (com exceção de 50 milhas)

ALVINN

- Utiliza uma Rede Neural
  - 960 entradas
    - Matriz 30x32 derivada dos *pixels* de uma imagem
  - 4 unidades intermediárias
  - 30 unidades de saída
    - Cada uma representando um comando para a direção

Carros da Google

- Stanford Artificial Intelligence Laboratory
  - Sebastian Thrun
- Comunicação por sensor (topo do carro)
  - Recebe informação do Google street view
  - Atua no volante de direção e nos pneus
  - 175,000 milhas sem acidentes
- Estado de Nevada aprovou lei permitindo driverless cars (Março 2012)

André de Carvalho 13

Google car

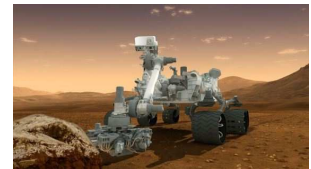
http://www.omg-facts.com/Technology/Google-has-developed-a-driverless-car/51099

André de Carvalho 14

Google car

André de Carvalho 15

Curiosity

- Robô Mars
- NASA e Jet Propulsion laboratory
- Mais de 1 tonelada

André de Carvalho 16

Monotrilho Palm

- Dubai

André de Carvalho 17

Algoritmos de AM

- Grande número
  - Agrupamento de dados (K-médias)
  - Algoritmos de indução de Árvores de Decisão
  - K-NN
  - Máquinas de Vetores de Suporte
  - Naive Bayes
  - Raciocínio Baseado em Casos
  - Redes Neurais Artificiais
  - Sistemas Inteligentes Híbridos

André Ponce de Leon F de Carvalho 18

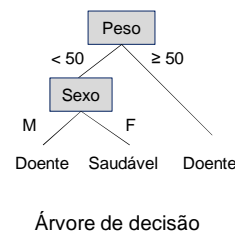
Algoritmos de AM

- Podem ser agrupados por diferentes critérios
  - Baseados em distâncias
    - K-NN
  - Baseados em otimização
    - RNs
  - Baseados em probabilidade
    - NB, SVMs
  - Baseadas em procura
    - Indução de ADs

Viés indutivo

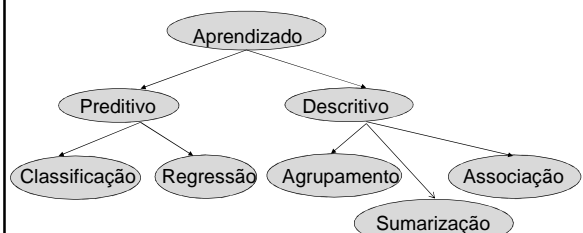
- Indução de hipóteses
  - Aprender a partir de um conjunto de exemplos
    - Induzir modelo ou hipótese
      - Aplicar a novos dados
- Todo algoritmo de AM indutivo tem um viés
  - Tendência a privilegiar uma dada hipótese ou um dado conjunto de hipóteses

Viés indutivo

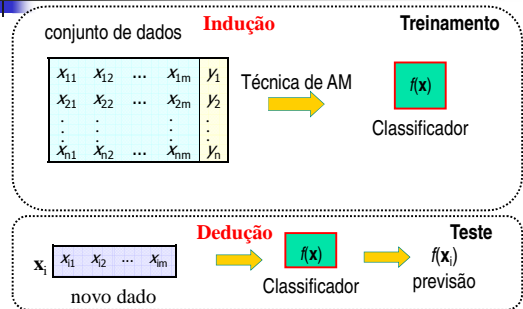
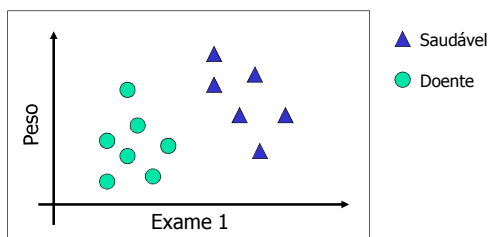
- Pode ser:
  - Viés de preferência ou busca
    - Como as hipóteses são pesquisadas no espaço de hipóteses
    - Preferência de algumas hipóteses sobre outras
    - Ex.: preferência por hipóteses simples (curtas)
  - Viés de representação ou linguagem
    - Define o espaço de busca ou de hipóteses
    - Restrição das hipóteses que podem ser geradas
    - Ex.: hipóteses podem conter apenas regras conjuntivas

Viés de representação

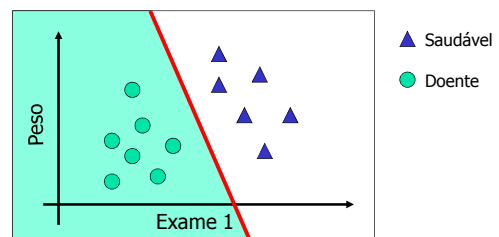
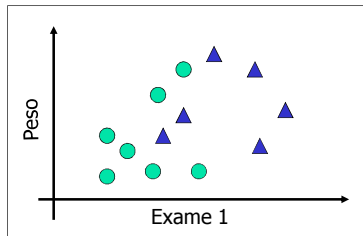
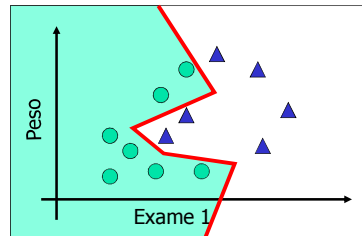
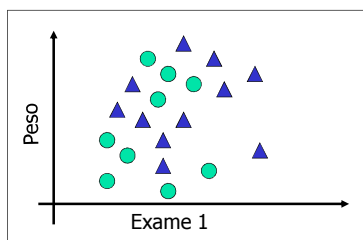
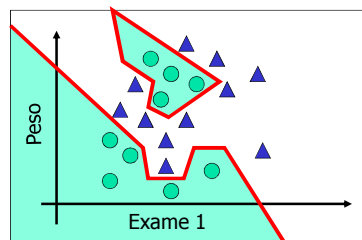
Viés indutivo

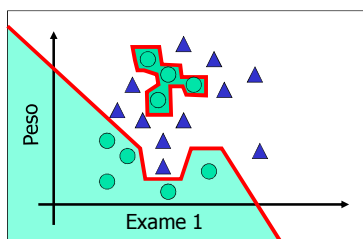
- Algoritmos de AM precisam ter um viés indutivo
  - Necessário para restringir o espaço de busca
  - Se não houvesse viés não haveria generalização
    - Regras / equações seriam especializados para os exemplos individuais

Tarefas de aprendizado

Conjunto de Dados
Atributos de entrada (preditivos)
Nome Temp. Idade Peso Altura
João 37 70 94 190 Saudável
Maria 38 65 60 172 Doente
José 39 19 70 185 Doente
Sílvia 38 25 65 160 Saudável
Pedro 37 70 90 168 Doente
Exemplos (objetos, padrões)
Atributo alvo
25

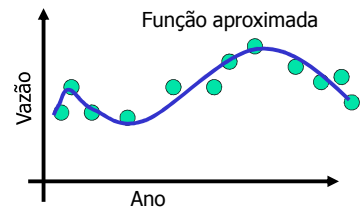
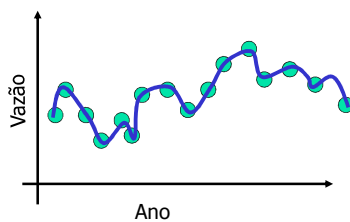
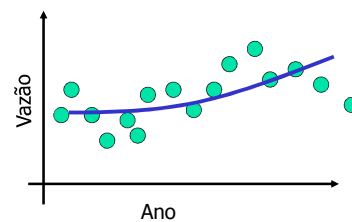
Métodos Preditivos
André Ponce de Leon F de Carvalho
26

Classificação
Objetivo: aprender uma função que mapeia um exemplo em uma dentre N classes
Exemplos:
Classificar aplicação para um cartão de crédito como boa ou ruim
Classificar tecido como normal ou cancerígeno
Definir se um paciente tem ou não uma doença
André Ponce de Leon F de Carvalho
27

Classificação
conjunto de dados
Indução
Treinamento
Técnica de AM
f(x)
Classificador
Dedução
Teste
novo dado
Classificador
previsão
André Ponce de Leon de Carvalho
28

Classificação
Como classificar?
Peso
Exame 1
Saudável
Doente
29

Classificação
Problema linear
Peso
Exame 1
Saudável
Doente
30

Classificação
Como classificar?
Peso
Exame 1
Saudável
Doente
31

Classificação
Problema não linear
Peso
Exame 1
Saudável
Doente
32

Classificação
Como classificar?
Peso
Exame 1
Saudável
Doente
33

Classificação
Problema não linear
Peso
Exame 1
Saudável
Doente
34

Classificação
Overfitting
Peso
Exame 1
Saudável
Doente
André C P L F de Carvalho
35

Classificação
Algoritmos
Árvores de Decisão (C4.5)
Conjuntos de regras
Redes Neurais
Máquinas de Vetores de Suporte
K-NN
Regressão Logística
Redes Bayesianas
André Ponce de Leon F de Carvalho
36

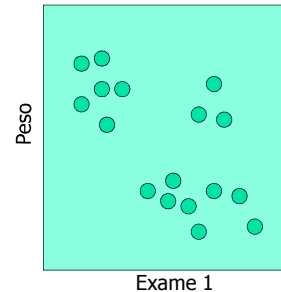
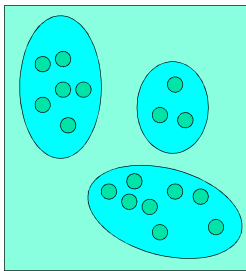
Regressão

- Objetivo: aprender uma função que mapeia um exemplo em um valor real
  - Caso especial: análise de séries temporais
- Exemplos:
  - Prever valor de mercado de um imóvel
  - Prever o lucro de um empréstimo bancário

Regressão

- Técnicas
  - Árvores de Regressão
  - Redes Neurais Artificiais
  - Máquinas de Vetores de Suporte
  - Regressão Linear

Agrupamento (Clustering)

- Objetivo: organizar exemplos não rotulados em grupos (clusters)
  - De acordo com uma medida de similaridade ou correlação entre eles
  - Aprendizado não supervisionado
- Não existe conhecimento anterior sobre:
  - Número de grupos (várias vezes)
  - Significado dos grupos

Agrupamento

Agrupamento

Agrupamento

- Técnicas
  - Redes Neurais SOM
  - K-médias
  - FCM
  - DBSCAN
  - Single-Link

Sumarização

- Objetivo: encontrar descrição simples e compacta para um conjunto de dados
- Freqüentemente utilizada para:
  - Exploração interativa de dados
  - Geração automática de relatórios
  - Exemplo:
    - Definir o valor médio de compras feitas nos finais de semana em um supermercado

Sumarização

Nome	Idade	Sexo	Altura	Tem filhos
João	32	M	180	S
Maria	30	F	-----	N
Pedro	23	M	160	S
José	45	M	170	S
Sueli	18	F	175	N

Sumarização
Nome Idade Sexo Altura Tem filhos
João 32 M 180 S
Maria 30 F ----- N
Pedro 23 M 160 S
José 45 M 170 S
Sueli 18 F 175 N
Idade média: 29.6
Mediana da idade: 30
Sexo mais presente: M
Menor altura: 160
André Ponce de Leon F de Carvalho 49

Sumarização
Técnicas podem ser divididas em:
Simples:
Média
Mediana
Desvio padrão
Mais sofisticadas:
Regras de sumarização
Técnicas de visualização multivariadas
André Ponce de Leon F de Carvalho 50

Exercício
Sumarizar cadastro de pacientes abaixo:
Nome Febre Enjôo Manchas Dores Diagnóstico
João sim sim pequenas sim doente
Pedro não não grandes não saudável
Maria sim sim pequenas não saudável
José sim não grandes sim doente
Ana sim não pequenas sim saudável
Leila não não grandes sim doente
Luis não sim grandes sim doente
Lívia não não pequenas sim saudável
André Ponce de Leon F de Carvalho 51

Regras de Associação
Objetivo: dado um conjunto de itens e uma base de dados de transações
Encontrar um conjunto de regras de associação entre os itens
Exemplo:
Procurar por itens que são freqüentemente comprados juntos
André Ponce de Leon F de Carvalho 52

Regras de Associação
Transação Itens comprados
1 pão, queijo, manteiga, massa
2 pão, geléia, suco
3 queijo, arroz, massa
4 massa, queijo
5 massa, queijo, pão
André Ponce de Leon F de Carvalho 53

Regras de Associação
Transação Itens comprados
1 pão, queijo, manteiga, massa
2 pão, geléia, suco
3 queijo, arroz, massa
4 queijo, vinho
5 massa, queijo, pão
40% dos clientes compram pão e queijo
75% dos clientes que compram queijo também compram massa
André Ponce de Leon F de Carvalho 54

Conclusão

- Mineração de Dados
- Aprendizado de Máquina
- Algoritmos
  - Viés indutivo
- Tarefas
  - Preditivas
  - Descritivas

André Ponce de Leon F de Carvalho 55